Data Exploration of Sentence Structures and Embellishments in German texts: Comparing Children's Writing vs Literature

Rémi Lavalley
Cooperative State University
Karlsruhe
Germany
lavalley@dhbw-karlsruhe.de

Kay Berkling
Cooperative State University
Karlsruhe
Germany
berkling@dhbw-karlsruhe.de

Abstract

It is of interest to study sentence construction for children's writing in order to understand grammatical errors and their influence on didactic decisions. For this purpose, this paper analyses sentence structures for various age groups of children's writings in contrast to text taken from children's and youth literature. While valency differs little between text type and age group, sentence embellishments show some differences. Both use of adjectives and adverbs increase with age and book levels. Furthermore books show a larger use thereof. This work presents one of the steps in a larger ongoing effort to understand children's writing and reading competences at word and sentence level. The need to look at variable from non-variable features of sentence structures separately in order to find distinctive features has been an important finding.

1 Introduction

Reading and writing are core competencies for success in any society. In Germany, the *Program for International Student Assessment* (PISA) study and the *Progress in International Reading Literacy Study* (PIRLS) (Bos, 2004) have shown that around 25% of German school children do not reach the minimal competence level necessary to function effectively in society by the age of 15. While the average performance is on par with other OECD countries, Germany falls short on higher levels of achievement and demonstrates a growing heterogeneity between genders and social backgrounds (Prenzel et al., 2013). Analyzing the types of errors that children make in their texts (Berkling and Reichel, 2014) it has been found that in the upper grades many grammatical issues persist that may be an indicator for the problems that become apparent in the above studies. It is therefore important to understand progression in sentence difficulty and its impact on didactics. Looking at research on sentence difficulty and text leveling, extensive research has been published for the English language. There are a number of works on defining sentence complexity or readability (Glöckner et al., 2006), (DuBay, 2008), (Sitbon and Bellot, 2008), (Benjamin, 2012), (Nelson et al., 2012), (Vajjala and Meurers, 2014). Sentence length, adverbs, morphemes, lexical analysis are some of a large number features that are used. Very often these features however do not represent the order in which the words appear in the text. Only few authors look at sentence structure and parse tree architectures (Schwarm and Ostendorf, 2005). In contrast, for German very few studies on this subject can be cited (Bamberger and Vanecek, 1984), (Hancke et al., 2012). Classifiers use some of the same features that had been used for English to classify difficulty levels of texts into major categories (child vs. adult writing). However, at this time, an automated categorization of reading texts for German does not exist. While there are some rules on readability, these are not defined at fine

grained levels with a clear progression and therefore also not automated or tested in a systematic manner on readers.

Given the existing body of knowledge, it became clear that some fundamental research is needed in looking at the sentence construction in data before moving on to a discussion about difficulty levels. This paper therefore presents a systematic approach to automatically analyse existing texts. The first goal is to gain a deeper understanding of German sentence structure and its occurrence patterns in different types of written texts, namely children's literature and children's writing for beginners, fourth graders and eight graders.

After an Introduction, Section 2 will review the structure of the German sentence. Section 3 will detail the data that was used for the exploration. Section 4 and 5 describe the automatic processing of the data. Section 6 will present results. Section 7 draws conclusions for future work.

2 Parsing German Sentence Structure

In order to understand how sentences will be analyzed, this section will review German sentence structures, verb valency and adjective and adverbial embellishments.

2.1 Features Description

The following list denotes the German standard sentence structures:

V2: This is the most common structure in German language. The verb is in second position. The subject can be either in first position (*Er* ***arbeitet*** *viel. He* ***works*** *a lot.*), or placed after the verb if the first position is used by something else, such as an adverb (*Jetzt* ***arbeitet*** *er. He* ***is working*** *now.*) or interrogative word (*Wo arbeitet er? Where does he work?*). Some words are not counted in order to determine the verb position. For instance, coordinate conjunctions (*Und er* ***arbeitet***. *And he* ***works***). In this case, the verb is considered to be in second position.

V1: The verb is in first position. This structure is generally used for the imperative form (*Sei ruhig! Be quiet!*) or for the interrogative form without interrogative word (*Hast du Hunger? Are you hungry?*)

VE: The verb is in final position. Generally used in subordinate clauses (*Ich denke,* ***dass*** *er zu viel arbeitet. I think* ***that*** *he is working too much.*)

NV: nominal clauses.

2.2 Valency

Verb valency refers to the number of arguments required by a verbal predicate. It includes the subject as well as the objects of the verb. (Ágel and Fischer, 2010). For the purpose of this work, only the items that have actually been attached are considered. The following example demonstrates this: *Lola gave her book.*, vs. *Lola gave her book to Lio.*. While the maximum number of arguments (theoretical valency) for the verb *to give* is 3: Subject (Lola), direct object (her book), indirect object (Lio), in the first sentence the 'used valency' is 2.

2.3 Adverbs and Adjectives

Carefully used adjectives and adverbs can be an indicator of a writers' command of the language. For example, *The little dark blue Smurf with glasses is really embarrassed* can be considered more descriptive writing than simply *The Smurf is embarrassed*. The study of adverbials complements that of verb valency. Consider for instance the two following sentences: *Ich gehe in die Schule. - I go to the school.* vs. *Jetzt gehe ich in die Schule. - I go to the school now.*. In both cases Valency equals 2 (*Ich - I*, *in die Schule - to the school*), while the feature of counting adverbs provides additional information about the usage of a temporal adverbial as *embellishment* to the original sentence construction.

3 DATA

The data chosen for this study comes from the Karlsruhe Database of children's writing (Berkling et al., 2014) and a selection of children's books.

3.1 Texts for Children (Books)

The corpus of literature was obtained through a random selection of books that are commonly read (as defined by the local public library) by children at the selected age groups[1]. Only Ger-

[1]**Grades 1 and 2:** Ages tend to be between 6 and 8 (merged into Grade 2); **Grade 4** Ages tend to be between

Grade No.	2	4	8
# books	21	15	11
Sentence length	7.4	9.7	12.1
# sentences kept	935	869	797
# children texts	237	258	245
Sentence length	10.5	13.3	12.5
# sentences kept	869	2133	1698

Table 1: Number of texts, average sentence length and sentences kept per grade, for both corpora

man authors were selected to eliminate effects of translation on quality. From each book, sample pages were selected and digitized resulting in the copurs statistics given in Table 1.

3.2 Text by Children (Childrens' Writings)

The children's data was collected in 2011–2013 from elementary schools and two types of secondary schools, Realschule and Hauptschule. Students' text was elicited in order to obtain an extended amount of freely written texts. The collection includes 1,752 texts from 1,730 students from grade 1 through 8 and is described in detail in a corresponding publication (Berkling et al., 2014).

The data is transcribed both in its original form (with spelling errors) and in a corrected version called target. While the target sentence has correctly written words, the grammatical errors and erroneous sentence structures remain leading to a non-trivial task of sentence structure analysis. For this study a subset of 740 texts written by children from grades 1, 2, 4 and 8 have been considered. The general statistics are summarized in Table 1.

4 Data Preparation

4.1 The Parser

All sentences in the databases were automatically parsed using the Berkeley's parser (Petrov et al., 2006) for German, with *-tokenize* (to use the integrated tokenizer) and *-accurate* (favours accuracy over speed) options. An example of such a parsing looks as follows, for the sentence *Das gibt ein Durcheinander!* (*This is a mess!*):

Output: *((PSEUDO (S (PDS Das) (VVFIN gibt) (NP (ART ein) (NN Durcheinander))) ($. !)))*

4.2 Sentence Decomposition

Given the parser output, a tool was developed to automatically classify the structure of the sentences. While finding the different clauses is generally done by the parser, a few manual rules to overcome the parser errors were added. The tool isolates the different components of a clause (POSTAG word) and stores them in a table in order of occurrence. Some components are thus grouped with higher entity, while others are not: In the example given above: *gibt* is tagged independently (VVFIN gibt) and in (NP (ART ein) (NN Durcheinander)) the parser has recognized a noun-phrase *ein Durcheinander* (*a mess*) and provides information about the different words (article and noun). We considered the external component as an entry in our table. Except in case of Verb phrase (VP), where the tool doesn't consider VP as one component but uses the isolated words information, as for instance, with this sentence:
(PDS Das) (VAFIN hat) (VP (PPER Karolina) (PRF sich) (AVP (ADV schon) (ADV immer)) (VVPP gewünscht)) [...]
In this case it's interesting to have the components of the VP as different entries. To compute the Valency (see Section 2.2), we need to additionally extract *Karolina* (Subject of the verb).

4.3 Data Cleaning

Some sentences were removed from both of the corpora, if they were too short (less than three words) or too long (more than 50 words). These lengths generally resulted from errors in the previous steps, such as transcription or OCR errors (e.g., a missing dot that leads to very long sentences). Analysing the data in a first path resulted in a very large number of different combinations of sentence structures. Given that most types occurred only in few sentences, the analysis will concentrate only on the 22 different structures that occur at least ten times on both of the corpora. Table 1 shows the number of sentences kept for the rest of the work presented here.

9 and 11; **Grade 8 and 8+:** Ages in this grade vary around 14 (merged into Grade 8)

5 Sentence Analysis

5.1 Sentence Structure

A clause structure is determined by the position of the main verb (finite) in the clause, which is tagged as V*FIN (VVFIN, VAFIN for auxiliaries, VMFIN for modal verbs). The tool categorizes a clause according to the structures defined in Section 2.1 by looking for verbs in their position.

5.2 Complex Structure Recognition

Many sentences consist of several clauses. The representation of the entire sentence therefore consists of a combination of classified clauses. The tool thus tags the entire sentence with the following notation scheme for Coordinate Clauses **CC** and Subordinate Clauses **SC** as examplified below.

CC: V2-V2 Ich$_{pos=1}$ mag$_{pos=2}$ das$_{pos=3}$ nicht$_{pos=4}$, aber$_{pos=0}$ ich$_{pos=1}$ gehe$_{pos=2}$ mit$_{pos=3...}$ ihnen ins Kino. *I don't like this, but I go to the cinema with them*.

SC: V2[VE] The verb is in second position in the main clause and in the final position in the subordinate clause. *Ich denke, dass ich ins Kino gehen* werde. *I think that I* will *go to the cinema.* In this case, an auxiliary verb is used in the subordinate clause to build the future tense (werde/will), this one is the conjugated verb and stands at the end of the clause.

SC: V2[VE]# The sharp symbol (#) is used to denote the fact that the subordinate clause occurs before the main clause. *Wenn du mir ein Blatt Papier gibst,* schreibe *ich dir einen Brief. (If you give me a paper, I* write *you a letter.*)

SC: V2[V2]# In this structure there is a main clause and a subordinate one, the subordinate stands before the main clause and they both have verbs in second position. In our corpus, it's mainly related to dialogs (*Ich rufe dich an,* sagt *Lola. I call you,* says *Lola.*). In this example, the subordinate clause is *Ich rufe dich an* (it is what Lola says) and in the main clause, the verb is considered as in second position because the first position is occupied by the subordinate clause.

There can be more than two clauses, such as 3 coordinates, one main clause with two interlocked subordinates (V2[NV[VE]]), one main clause with a subordinate made of two coordinates clauses (V2[VE-VE]), to name a few. The tool can represent all of these combinations.

5.3 Evaluation of Structure Classification

The tool developed for sentence structure analysis has been evaluated on 400 sentences manually annotated: 200 sentences coming from books and 200 from children writings. We also annotated these sentences as correct or not. 20% of the sentences extracted from books contained errors introduced during digitization: Non-existing words, space missing or added, or punctuation marks missing, sentences erroneously merged into one, missing comas, making it difficult to determine clauses (in German, subclauses are separated by commas). 30% of the sentences in children's writing contained errors: Spelling errors not corrected by annotators, grammar errors, such as words in wrong position, usage of an incorrect word (not corrected by annotators), sentences intentionally concatenated into one by the writers (making them difficult to parse). The overall precision of the tool is the same for both corpora: 80% of the sentences correctly labeled. The system has wrongly labeled structures for 38 sentences of the Books corpus (16 of these sentences had at least 3 clauses) and 41 of the children corpus (27 had at least 3 clauses). More than 3 clauses are usually a sign of bad sentence construction and are therefore difficult to parse.

5.4 Valency

The tool computes the number of arguments by going through the table containing the constituents of the sentence. Constituents are counted towards the valency counter as long as they are not excluded given the rules below. These are intended to bypass parser mistakes while keeping it as exhaustive and accurate as possible.

Word: The list denotes a number of POS-tags that cannot be a subject or an object of a verb, such as articles, other verbs (infinite, participles), preposition, adjectives, adverbs, and particles (separable verbs). If a word is not labeled with one of these POS-tags (KOUS, PTK, ADV, KON, V* denoting different types of verbs...) then it is an argument of the verb.

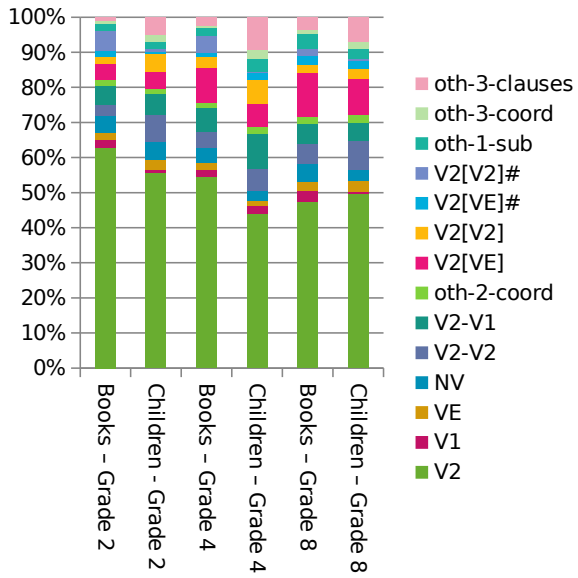

Figure 1: Partition of sentence structures, for both corpora at different grade levels.

Clause: Some clauses can be the object of a verb: *Ich denke, dass er zu viel arbeitet*. *Ich* (*I*) is the subject, while *dass er zu viel arbeitet* (*that he works too much*) is object of the verb *denke* (*think*). Other clauses cannot be objects. For example, *Er ist mehr intelligent, als ich. (He is more intelligent than me)*. These clauses are excluded by using a list of POS-tags (KOUS, KON, PROAV, KOKOM) combined with the list of words that introduce adverbial clauses (”bevor”, ”als”, ”wenn”, ”während”, ”indem”, ”solange”, ”bis”, ”weil”, ”da”, ”wie”, ”damit”, ”obwohl”, ”trotzdem”, ”obgleich”, ”denn”, ”seitdem”).

5.5 Adverbs and Adjectives per Sentence

Adjectives and adverbs are easily detected by the POS tags provided by the parser. The tool counts the number of words labeled as adverbs or adjectives according to the parser.

6 Results

6.1 Sentence Structures

In both corpora the average occurrence frequency of sentence structure per grade was computed as well as average use of adjectives/adverbs per sentence type for each sentence type and grade level. Figure 1 shows the partition of sentences structures by grade and corpora. Less frequent structures were merged into one of four supercategories: **”oth-2-coord”** contains all the sentences made of two coordinates clauses except V2-V1 and V2-V2, **”oth-1-sub”** contains sentences with a main clause and a subordinate clause other than the ones provided separately, **”oth-3-coord”** contains the sentences made of three coordinate clauses and **”oth-3-clauses”** the sentences made of 3 clauses including at least one subordinate.

We can observe the following: From Grade 2 (including Grade 1) to Grade 8, books use a decreasing number of V2 sentences (from 62% to 48%). Meanwhile children always have more or less 50% of their sentences of type V2. Children write a larger number of coordinate clauses (V2-V2, V2-V1, other 2 coordinates and other 3 coordinates) when compared to books. Inspecting the data, it can be seen that children create their own grammar rules and forget to split sentences. As children get older, they use less nominal sentences (NV). The proportion of subordinates clauses with verb ending (V2[VE]) increases with the grades in books (from 4 to 13%) - the same applies for children between 2nd and 8th grade. Children don’t really use inverted clauses (notation ending in #). In books these mainly occur with dialogues (*Ich arbeite nicht, sagt Lola - I don’t work, says Lola*), whereas the topics on which the children had to write didn’t especially involve dialogues even if some can be found in the texts. When children reach 8th grade, they tend to use the same structures that occur in books, i.e. the distribution of sentence types is roughly the same as that of 8th grade published literature.

6.2 Adverbs and Adjectives

Figure 2 depicts the mean number of adverbs used in sentences by books and children for the different structures. The last column is the mean number of adverbs on all the sentences, regardless of their structure. We can observe that children use almost as many adverbs as authors of books. The gap between Grade 2 and Grade 8 is more significant in books’ texts than in childrens’ writings. However, half of sentences have no adverbs at all (in Grade 2, it concerns 53% of children’ sentences and 54% of the books’ ones). This frac-

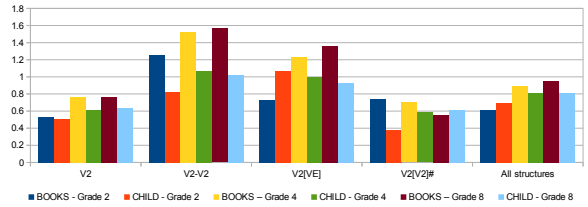

Figure 2: Mean number of adverbs per sentence for selected structures, for both of the corpora (Books or Children) at the different grades (2, 4, 8).

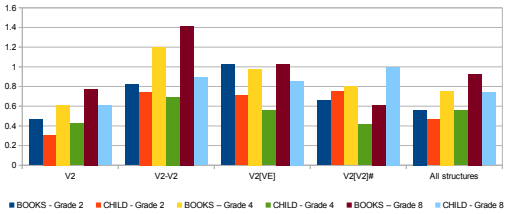

Figure 3: Mean number of adjectives per sentence for selected structures, for both of the corpora (Books or Children) at the different grades (2, 4, 8).

tion decreases with higher grades for both children (48% in 8th grade) and books (45%). Generally, the children have more sentences without adverbs than authors. The same observations can be made regarding the adjectives: in 2nd grade, 58% of books' sentences and 63% of children's ones don't have adjectives at all, whereas in 8th grade, these ratios are respectively 45% and 48%.

Figure 3 depicts the mean number of adjectives per sentence for different sentence structures. Such as for adverbs, children use them a little bit less than authors do, but more and more as they are getting older. The mean number of adjectives increased by 50% between Grade 2 and Grade 8.

6.3 Valency

According to our analysis, valency seems roughly invariant to age and the two text corpora used. The only kind of sentences on which significant differences have been observed is the V2 type. As shown in Table 2, children in Grade 2 have a different usage of objects compared to books: 44% of their verbs have only one complement (i.e., generally the subject), while this proportion is only 31% in books. Whereas this number slightly decreases in books to reach 22% in 8th grade, the

	Books		Children	
	G 2	G 8	G 2	G 8
Val=1	31%	22%	44%	25%
Val=2	55%	58%	45%	50%
Val=3	12%	17%	9%	19%
Val=4	2%	4%	0%	4%

Table 2: Proportion of V2 sentences having Valency = 1 to 4, for both of the corpora, at grades 2 and 8.

children use really less constructions of this type compared to their early ages to reach 25% of their sentences, which is close to the proportion observed in books. Accordingly, the global repartition between the different valencies of verbs is the same for books and children's writings when they reach 8th grade.

7 Conclusion and Future work

The goal of this work is a systematic approach to automatically analyze large amounts of texts and their structures to gain a deeper understanding on tackling text difficulty. Rules to recognize typical German sentence structures were implemented based on the output of an open source POS-tagger. Looking at texts written by and for children, the sentences were analyzed based on the occurrence distribution of particular structures within the texts at different grade levels. In addition, embellishments clues (valency, adjectives and adverbs) were counted and compared in their mean occurrence within sentences. It was found that children in 2nd grade have a personal way of writing (e.g., structures used are different from those of authors), while in 8th grade they are to some extent getting closer to the level of writing of the books. Increasing use of adjectives and adverbs over the years approach the profiles found in literature. Future work includes looking at correlations of features and adding information about word usage, spelling errors and semantics. A significant gap between leisure reading and children's texts with respect to their textbooks is observable. Further study needs to quantify that and determine a reasonable progression for didactics to advance students' towards academic skills.

References

Vilmos Ágel and Klaus Fischer. 2010. Dependency Grammar and Valency Theory. *Bernd Heine & Heiko Narrog (Hgg.), The Oxford Handbook of Linguistic Analysis, Oxford*, pages 223–255.

Richard Bamberger and Erich Vanecek. 1984. Lesen-Verstehen-Lernen-Schreiben: Die Schwierigkeitstufen von Texten in deutscher Sprache. *Wien: Jugend und Volk*.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Kay Berkling and Uwe Reichel. 2014. Der phonologische Zugang zur Schrift im Deutschen.

Kay Berkling, Johanna Fay, Masood Ghayoomi, Katrin Heinz, Rémi Lavalley, Ludwig Linhuber, and Sebastian Stücker. 2014. A Database of Freely Written Texts of German School Students for the Purpose of Automatic Spelling Error Classification. In *LREC Conference*.

Wilfried Bos. 2004. IGLU: Einige Länger der BRD im nationalen und internationalen Vergleich.

William H. DuBay. 2008. The principles of readability. 2004. *Costa Mesa: Impact Information*, 76.

Ingo Glöckner, Sven Hartrumpf, Hermann Helbig, Johannes Leveling, and Rainer Osswald. 2006. An architecture for rating and controlling text readability. *Proceedings of KONVENS 2006*, pages 32–35.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability Classification for German using Lexical, Syntactic, and Morphological Features. In *COLING*, pages 1063–1080.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440.

Manfred Prenzel, Christine Sälzer, Eckhard Klieme, and Olaf Köller. 2013. PISA 2012: Fortschritte und Herausforderungen in Deutschland. *Münster: Waxmann*.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530.

Laurianne Sitbon and Patrice Bellot. 2008. A readability measure for an information retrieval process adapted to dyslexics. In *Second international workshop on Adaptive Information Retrieval (AIR 2008 in conjunction with IIiX 2008)*, pages 52–57.

Sowmya Vajjala and Detmar Meurers. 2014. Exploring Measures of "Readability" for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 21–29.